Report of the Bioinformatics Working Group of the National Advisory Research Resources Council

June, 2000

Bioinformatics Working Group Goals

In 1999, NIH Director Harold Varmus appointed an advisory group called the Working Group on Biomedical Computing. The group produced a report recommending that the NIH undertake an initiative called the Biomedical Information Science and Technology Initiative (BISTI). That report is henceforth called the BISTI Report.

In May, 2000, National Center for Research Resources (NCRR) Director Judith Vaitukaitis appointed a National Advisory Research Resources Council Bioinformatics Working Group (BWG). Its goals were to relate the BISTI Report recommendations to the NCRR mission, and to recommend specific actions, priorities, and implementation strategies for the BISTI to NCRR.

Terminology

The BISTI Report uses terms such as biomedical computing, computational biology, and bioinformatics. The BWG noted that precise, universally accepted definitions of these terms are not available. Different scientists use these terms in different ways. Here is a summary of some of their common interpretations. Rather than treating what follows as a list of definitions, consider them to be possible meanings that different authors who use these terms may intend.

In the remainder of this report we use these terms interchangeably -- use of one term or the other in this report does not imply some fine gradation in meaning.

Biomedical Computing:
1. The application and development of computer methods for biomedical research.

Computational Biology:
2. A subfield of biology involving the computer analysis of biological data.
3. The application and development of computer methods for biological research.
4. The application and development of mathematical and algorithmic computer methods for biological research.

Bioinformatics:
5. The application and development of computer methods for biological research.
6. The application and development of computer methods for genomics and molecular biology.
7. The application and development of database-related computer methods for biological research.

Notes: (3) and (5) are intended to be identical.

Recommendations of BISTI Report

The original BISTI Report contained four recommendations:

- Establish between five and twenty National Programs of Excellence (NPEs) in Biomedical Computing.
- Establish a new program directed toward the principles and practice of information storage, curation, analysis, and retrieval (ISCAR).
- Provide resources for basic research (via R01s) to adequately support biomedical computing.
- Foster a scalable national computer infrastructure to provide biomedical researchers the computing resources they need.

Working Group Reactions

The BWG supported the BISTI Report enthusiastically. The report accurately identifies important potential contributions of the field of biomedical computing, such as contributing to a new system-level, integrative paradigm of biomedical research. It also identifies important challenges facing the bioinformatics field, and its recommendations will go a long way toward advancing the field.

The BWG felt that the NCRR is well positioned to implement many aspects of the BISTI, for example because the National Programs of Excellence proposed by the BISTI Report correspond closely to NCRR Resource Centers. The group also supported the BISTI Report recommendations for strong support for education because of the shortage of qualified bioinformatics personnel. The group recommended that NCRR strike an appropriate balance between new National Programs of Excellence, and new R01-type research grants.

The BISTI is long overdue. There has been a significant imbalance between the NIH investments in high-throughput data generation projects and in bioinformatics. It is ironic that the BISTI will begin the year that the first draft of the human genome is completed.

National Programs of Excellence in Biomedical Computing

The group recognized that there are at least two approaches to focusing national programs of excellence: first, on a computational problem area, such as simulation, molecular modeling, database technology, biomedical imaging and processing, or visualization; second, on a biological research area, such as molecular cell biology, genomics, image-guided surgery, neurobiology, or structural biology.

Issues that should be addressed by NCRR in considering NPEs for funding include the following: What is the mechanism for transfer of technologies from one center to another, and to the larger biomedical community? How will connections to DOE and NSF centers be established and maintained? Will computer equipment be adequately funded not only for purchase but also for maintenance and update (given the rapid obsolescence of computers)? Each NPE should balance the following components: research, software and database development, training, service, new methods for dissemination of software resources, and support of bench scientists. Different centers will emphasize these components in different ways depending on their expertise and strategic goals.

NPEs should vary in size from small multi-investigator groups to large centers. In addition to fostering synergistic interactions between a critical mass of experts, large centers will be needed to develop the large and complex software systems and databases that are required for impact in this field. Just as large-scale genomics projects are needed to generate large quantities of high-quality data, large-scale

software-development projects are needed to develop bioinformatics applications that can manage, disseminate, and analyze that data.

The group suggests that the NPEs be organized within a loosely interacting network to foster communication and exchange of computing technologies. In this network setting, some NPEs may support other NPEs by providing services such as a software clearinghouse, support for hardening software, and/or support for data interoperability and ontologies.

Given that hardware and staff support are provided for the NPE sites, they will be ideal facilities to carry on educational components that are badly needed for the biological users, and also for training a new generation of computational biologists. It is imperative that the NPE periodically sponsor workshops, symposia and distant-learning tutorials so that a broad community will be trained to use existing bioinformatics tools, and be stimulated to contribute to the development of new tools.

Information Storage, Curation, Analysis and Retrieval

The ISCAR area can be divided into two subareas: creation of new content databases, and development of new software for database management and analysis. The BWG felt that both areas are of great importance.

The group also felt that ISCAR is an area of bioinformatics where dramatic gains are possible, for several reasons. First, databases are important in virtually every biological area of bioinformatics. Second, the injection of relatively small amounts of software expertise into existing database projects will result in large gains, such as in the area of database interoperability. Third, propagation of errors and of unfinished data within the public databases is having a significant negative impact on their utility. The BWG noted that there is a particularly acute shortage of qualified personnel in this area.

The BWG noted that database content projects need special consideration in the NIH review process since they are resources, not hypothesis-driven research, yet are sometimes rejected by certain NIH review panels specifically because they are not hypothesis-driven research.

The BWG recommended that the NIH develop higher standards for database content projects. Projects should (a) develop an interoperability plan from inception, (b) employ a database management system (as opposed to flatfiles) when appropriate, (c) better handle data provenance (tracking of data source), (d) develop better tools for ensuring data quality. The BWG also requested that NIH work with the bioinformatics community to develop long-term funding model(s) for public databases in addition to complete reliance on government funding.

The BWG identified several promising ISCAR-related research areas in addition to those discussed in the BISTI Report. (1) *Refreshing derived databases when the source databases change.* Many database projects extract a subset of data from one or more larger databases, then further curate it based on the investigators' specific expertise. No automated method exists to merge later updates to the source database with changes made to the extracted data, and in many cases the source of the derived data is not even tracked. Consequently, merging of updates from the several sources usually does not occur. This leads to a morass of databases whose content overlaps but is inconsistent. (2) *Development of shared ontologies (database schemas, or data models).* Sharing ontologies can speed database development, reduce their semantic heterogeneity, and simplify the database interoperation problem. (3) *Empirical studies of bioinformatics databases as large and complex artifacts.* These databases integrate data from many sources, and transform that data using a mixture of computer programs and manual annotation. Thus they become so complex as to require systematic empirical study, such as to measure the error rate

within a database, and to verify whether the database is internally consistent.

Large Hardware Resources

There are not enough CPU cycles for biomedical researchers. The NIH should support, in concert with NSF, localized super computing based on scalable clusters of commodity processors. The goal should be to provide scalable clusters, data servers, and high-speed network access in every biology department. Scalable clusters require significant expertise for their configuration and operation. Therefore, NIH should establish a clearinghouse for support expertise and resource scheduling software. Moreover, some important large projects require resources well beyond the capabilities of what is likely to be found in most biology departments. NIH should support large centralized resources, built on scaled-up versions of the same commodity processors, and with robust software and experienced staff, to enable such projects to be carried out in a timely fashion.

Production-Grade Software

An important issue for NIH will become the conversion of academic software to production-grade. The issues involved are not at all clear-cut, but have profound implications for bioinformatics software usability. We believe this is an important and complicated issue that should be studied more thoroughly.

Software becomes production-grade when it can be installed at external sites; operates on multiple platforms; has documentation; has been quality assurance tested; and is provided in conjunction with user support services for answering questions, fixing bugs, and delivering enhancements on an ongoing basis. Industry-wide, approximately 20% of software engineering effort goes into producing the first working code that implements a particular software artifact, and approximately 80% goes into making it production-grade and maintaining it as such. Two ways to achieve production-grade software are through commercialization or through academic production-grade efforts.

Advantages of commercialization include: The expense of software professionals can be amortized by economies of scale across multiple projects; software companies usually have a permanent help desk and bug reporting/tracking/fixing facilities; the development personnel may have less turnover than in the fluid academic environment; quality assurance and testing may be more systematic and thorough; porting the software to other platforms may be easier; the software may be integrated into a seamless package with other software products; the user interface may be more systematic and standard.

Disadvantages of commercialization include: software can stagnate due to lack of scientific expertise within the company; software can become unavailable to academics if a company decides that potential profits lie in the commercial market rather than the academic market; the company may see more opportunity for profit in other products and allow the software to languish; source code may become proprietary and thus eliminate opportunities to add new features; companies may go bankrupt and change ownership, again restricting software availability.

Advantages of academic production-grade efforts include: The developers understand both the scientific goals and the software; the scientific and mathematical expertise of the developers may be greater, resulting in a more effective program; having the original developers do the conversion leads to greater conceptual integrity in the final product; academic developers may have better access to working scientists as alpha and beta testers, leading to better customization to its intended task; academic developers may be more likely to make the software source code available to others.

Disadvantages of academic production-grade efforts include: When government funding ends, support and availability of the software may end; new versions of computer languages, data standards, and operating systems may make the software obsolete; the academic developers may lose interest in the software once all research issues have been solved; non-standard coding practices ("kludges") may appear more often; bug fixes and upgrades may not be done; the software may never reach production-grade at all if the developers have no expertise in production software; testing and quality assurance may be inadequate.

The expense of a given "hardening" project can be significant, and should be justified by high usage and value. But in many cases, the BWG expects significant value will be obtained by hardening efforts, and it is important to realize the full potential of the past NIH investment in a particular software project by elevating the software to production grade for widespread usage. One mechanism for funding hardening projects is for investigators to apply for supplements to existing grants to deliver production quality software.

Open-Source Software Distribution

"Open source" software distribution means distributing software source code along with executables. This model allows other scientists to improve and verify the software. The open-source model should be encouraged, but not required, at all levels by NIH funding mechanisms.

Building the Biomedical Computing Community

NPEs should be designed to integrate the larger scientific community. NIH should encourage submission of investigator-initiated proposals allied to NPEs. The NPE funding mechanism should be designed to foster the development of Collaboratory software. Funding criteria should include university support for multidisciplinary research and education. The NPEs should be designed to encourage involvement of larger biomedical and computational communities.

Relationship to Computer Science

Successful execution of the BISTI will require partnerships between biologists and computer scientists. Software development is a difficult and complex undertaking to which the field of computer science brings many different methods and techniques. In many cases, existing computer-science methods are appropriate for solving problems in biomedical computation. Computer scientists trained at the bachelors and masters levels generally have the expertise to build software systems using existing software techniques. The larger and more complex a bioinformatics software-engineering project is, the more critical it is to involve bachelors and masters-level computer scientists in the development of that system.

A project that relies solely on biologists with little software training risks producing a software system that does not work, does not scale to larger problems, or that cannot evolve over time to meet the changing needs of its users. A significant fraction of software development costs occur in the maintenance and evolution phases, rather than in the initial development phase. One mark of professional software engineering is a large and complex software system that has an internal simplicity that facilitates later evolution of the code.

Many biomedical-computing problems are so complex that they require the development of new computational methods, which by definition includes a component of computer-science research. Generally, PhD-level computer-science training is appropriate for such problems. PhD-level computer

scientists are familiar with all the methods and techniques that masters-level graduates are exposed to, but also have much deeper training in a particular subfield of computer science. That training includes techniques for developing new algorithms, for analyzing those algorithms theoretically and empirically, for comparing new algorithms to existing methods, and for publishing algorithms in a manner that allows other researchers to recreate them.

Biomedical-computing problems have already led to new computer-science advances in areas such as database interoperation and machine learning. The challenging nature of bioinformatics problems, coupled with the easy availability of large datasets that can provide real-world problems to computer-science researchers, suggest that bioinformatics problems will continue to drive new advances in computer science.

Opportunities for Impact

The BWG recognized that the development of new bioinformatics technology infrastructure has the potential to impact many areas of biomedical research. Opportunities for impact include protein folding, drug design, neurological modeling, system-level biological simulation, membrane transport and signaling, whole-cell and organ-level modeling and imaging, rapid genome analysis and comparison, image analysis, digital infrastructure for biomedical knowledge and data, and biomedical decision support.

Members of the Bioinformatics Working Group

The BWG consisted of eleven researchers in biomedical computing, whose doctorate-level training ranged from biomedical research to computer science. The International Society for Computational Biology (ISCB) was involved in suggesting members for the BWG, and several of the BWG members were founders or board members of the ISCB.

Chair: Peter D. Karp, Ph.D.
Bioinformatics Research Group
Artificial Intelligence Center
SRI International
Menlo Park, CA

Wah Chiu, Ph.D.
National Center for Macromolecular Imaging
Baylor College of Medicine
Houston, TX

Susan Davidson, Ph.D.
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA

Mark H. Ellisman, Ph.D.
University of California San Diego
San Diego, CA

Terry Gaasterland

Laboratory of Computational Genomics
The Rockefeller University
New York, NY

Gwen A. Jacobs, Ph.D.
Department of Cell Biology and Neuroscience
Center for Computational Biology
Montana State University
Bozeman, MT

Ron Kikinis, M.D.
Department of Radiology
Brigham and Women's Hospital
Associate Professor of Radiology
Harvard Medical School,
Boston, MA

Teri Klein, Ph.D.
Stanford Biomedical Informatics Stanford University School of Medicine
Stanford, CA

Rick Lathrop, Ph.D.
Dept. of Information and Computer Sciences
University of California
Irvine, CA

Klaus Schulten, Ph.D.
Beckman Institute
University of Illinois
Urbana, IL

Gary D. Stormo, Ph.D.
Department of Genetics
Washington University Medical School
St. Louis, MO